

# pROST : A Smoothed $\ell_p$ -norm Robust Online Subspace Tracking Method for Realtime Background Subtraction in Video

Florian Seidel · Clemens Hage · Martin Kleinsteuber

March 29, 2013

**Abstract** An increasing number of methods for background subtraction use Robust PCA to identify sparse foreground objects. While many algorithms use the  $\ell_1$ -norm as a convex relaxation of the ideal sparsifying function, we approach the problem with a smoothed  $\ell_p$ -norm and present pROST, a method for robust online subspace tracking. The algorithm is based on alternating minimization on manifolds. Implemented on a graphics processing unit it achieves realtime performance. Experimental results on a state-of-the-art benchmark for background subtraction on real-world video data indicate that the method succeeds at a broad variety of background subtraction scenarios, and it outperforms competing approaches when video quality is deteriorated by camera jitter.

**Keywords** Background Subtraction · Robust PCA · Online Subspace Tracking · CUDA

## 1 Introduction

Many high-level computer vision tasks like object tracking, activity recognition and camera surveillance rely on a pixel-level segmentation of scenes into foreground and background as a preprocessing step. This task is often referred to as background subtraction and has drawn great attention in recent

years. Surveying the multitude of existing methods is beyond the scope of this article, and for this purpose we refer to two excellent recent surveys of the field, [8] and [14].

Robust Principal Component Analysis algorithms have been proven successful at separating foreground objects from a static or dynamic background [11]. The underlying assumption of Robust PCA is that the analyzed data can be considered a superposition of a low-rank and a sparse component, which can be separated blindly without any further assumptions on the data. For many video sequences this assumption holds true. The vectorized frames of a video background span a low-dimensional subspace, whereas rapidly moving objects appear sparse in space and time and thus can be distinguished from the background using Robust PCA. Most Robust PCA algorithms focus on processing the complete data set at once in a batch-processing manner. This means that all frames of the video sequence and their statistics are available and the algorithm performs background subtraction on the entire sequence. Recently, methods have been presented which allow for online subspace tracking [2], i.e. video data can be processed frame by frame and each new incoming data sample contributes to the estimate of the background.

This paper introduces a robust online background subtraction algorithm, called *pROST*: a smoothed  $\ell_p$ -norm Robust Online Subspace Tracking Method. The name reflects the two defining characteristics of the algorithm. Firstly, to achieve robustness against outliers we use a smoothed and weighted  $\ell_p$ -(pseudo)-norm cost function. Secondly, an efficient alternating online optimization framework for the estimating the subspace makes the algorithm suitable for online subspace tracking. The algorithm is tailored for real-time background subtraction in streaming video and makes use of the spatio-temporal dependencies between pixel labels, i.e. the foreground or background assignment on a pixel level. This leads to especially good performance in videos that require

Florian Seidel  
Department of Informatics  
Technische Universität München  
Boltzmannstr. 3, 85748 Garching, Germany  
E-mail: seidel.florian@gmail.com

Clemens Hage, Martin Kleinsteuber  
Department of Electrical Engineering and Information Technology  
Technische Universität München  
Arcisstr. 21, 80333 Munich, Germany  
E-mail: {hage,kleinsteuber}@tum.de

bootstrapping, which means learning a new background from corrupted data. It also alleviates problems with large foreground objects, which often arise in PCA-based methods [4]. In comparison to other methods we observe that our algorithm is particularly good at dealing with the varying background in video recorded by jittery cameras.

One of the main difficulties with comparing different background subtraction methods has been the lack of an accepted benchmark. Various data sets exist (e.g. [18] and [22]), which provide video sequences and manually segmented test images. However, the lack of pixel-level ground truth has led to rather selective evaluation instead of comparable and representative results, as the authors of the *SABS* dataset criticize [5]. They overcome the cumbersome task of hand-segmenting video sequences by providing an artificially rendered scene, which allows a very detailed and precise segmentation. Although the animations are close to photo-realistic, the visual impression is fundamentally different from true recordings.

In order to establish a benchmark on real-world video sequences, the *changedetection.net* dataset [10] has been introduced. It provides image sequences and full ground truth for a variety of categories such as static and dynamic background, thermal imaging and camera jitter, as well as the explicit distinction between foreground objects and their shadows. Seven statistical error measures are computed to evaluate the performance as detailed as possible. This prohibits tuning a method for a single performance measure and guarantees significant scores. Evaluation and thus the ranking of all competing methods are computed per category and as an overall average. All reported results are conveniently accessible on the project website <sup>1</sup>.

The paper is outlined as follows: in Section 2 of this paper we define our understanding of foreground and background, give a brief overview of the issues arising in background learning and maintenance and explain how a background model can be used for foreground segmentation. We describe how PCA can be used to create a model of the scene background and motivate the use of robust cost functions. In Section 3 we present the pROST framework, which is motivated and discussed in the context of background modeling and foreground segmentation. Section 4 provides details on the implementation on a graphics processing unit. We evaluate our algorithm on the *changedetection.net* dataset and discuss the results in Section 5. Concluding the paper, we analyze typical issues in the modeling of scene backgrounds with pROST and explain with a few examples how they are addressed by the choice of parameters.

## 2 Robust PCA based background models

Video background is commonly defined as the union of persistent elements of a scene. They can be static or may exhibit repetitive dynamics, which either occur on an object-level, e.g. an escalator or a fountain, or on a global scale, e.g. water or waving trees, but also camera jitter. In other words, the background is comprised of elements that are known, predictable and not of interest for higher-level tasks such as surveillance or activity recognition. Everything else that moves about in the scene is understood as foreground objects. From this definition the idea of treating foreground-background segmentation as an outlier detection problem arises naturally, i.e. a model of the background is established and the foreground is segmented by comparing each video frame to this model. Elements of the video frame that do not fit the background model are labeled as foreground, while the rest is labeled as background. Virtually every algorithm published so far follows this approach, but they differ in the type of background model that is used and how it is maintained, and in potential pre- and post-processing steps.

Establishing and maintaining an accurate background model is not trivial under real-world circumstances, and certain requirements have to be met by a method to be useful in challenging scenarios. For example, a scene cannot be observed in all possible lighting or weather conditions. Or it might be impossible to have a separate training stage in which the scene is free of foreground objects. Therefore, the ability to learn a background model from corrupted training data is of crucial importance. Without having any pixel semantics this is of course only possible if foreground objects have different statistical properties in time and space than the background. Furthermore, it is also necessary to update the background model continuously, which is commonly referred to as *model maintenance*.

### 2.1 PCA background models

PCA background models were introduced in [20] as part of a system for human activity recognition. The underlying assumption of PCA models is that a vectorized video background  $B \in \mathbb{R}^{m \times n}$  can be represented as a product of a subspace-defining matrix

$$U \in \text{St}_{k,m} := \{U \in \mathbb{R}^{m \times k} | U^\top U = I_k\} \quad (1)$$

and the subspace coordinates  $Y \in \mathbb{R}^{k \times n}$ .  $\text{St}_{k,m}$  denotes the Stiefel manifold and  $I_k$  is the  $k$ -dimensional identity matrix. With the supplement of a Gaussian noise matrix  $E \in \mathbb{R}^{m \times n}$  whose  $(i, j)$ -entries  $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma)$  are independent Gaussian random variables, this results in the data model

$$B = UY + E. \quad (2)$$

<sup>1</sup> [www.changedetection.net](http://www.changedetection.net)

Considering now a video sequence  $X \in \mathbb{R}^{m \times n}$ , which might also contain additional foreground objects, one can try to recover the matrix  $U$  that defines the subspace and  $Y$  as the solution of the optimization problem

$$\min_{U \in \text{St}_{k,m}, Y \in \mathbb{R}^{k \times n}} \|X - UY\|_2 \quad (3)$$

where  $\|\cdot\|_2$  denotes the  $\ell_2$ -norm. This is equivalent to the classic PCA problem [21], whose well-known closed-form minimizer is given by the leading  $k$  eigenvectors of the data covariance matrix.

Given a basis  $U$  for the background subspace and an observed image  $x \in \mathbb{R}^m$  a foreground segmentation mask  $F \in \{0, 1\}^m$  can be obtained through fitting the model by firstly solving

$$y^* = \arg \min_{y \in \mathbb{R}^k} \|x - Uy\|_2 \quad (4)$$

followed by applying a thresholding operation

$$F_i = \begin{cases} 1 & \text{if } |x_i - u_i^\top y^*| > \delta \\ 0 & \text{else,} \end{cases} \quad (5)$$

where  $u_i^\top$  denotes the  $i$ -th row of  $U$  and  $\delta > 0$  is a threshold parameter.

PCA-based methods for background estimation show great advantages over competing methods whenever backgrounds are dynamic and include illumination changes [4]. However, as the classical PCA approach uses the  $\ell_2$ -norm, the subspace reconstruction may severely suffer from outliers. In the practical context of background modeling, several problems can be observed: Firstly, in common PCA undue weight is given to the foreground elements when fitting the background model to camera frames during the segmentation process, which severely limits the admissible size of foreground objects. Secondly, images containing foreground objects can lead to corruption of the background model during bootstrapping and background maintenance. Finally, batch-processing results in tremendous memory requirements.

As discussed in [4], a lot of effort has been spent to overcome these limitations. The predominant mechanism for achieving robustness in PCA-based methods is weighting or replacing individual pixels in order to reduce the influence of known foreground objects on the background reconstruction error. Adaptive thresholding [24] has been proposed to allow for larger foreground objects and an attempt at a robust incremental estimation of backgrounds has been presented in [19].

## 2.2 Robust background estimation

The computer vision community has recognized that Robust PCA methods offer substantial advantages over classic PCA and background modeling has become increasingly

popular as an application of Robust PCA algorithms. For an overview of background subtraction using Robust PCA we refer to [11].

The typical assumption for Robust PCA [6] is a data model

$$X = L + S, \quad (6)$$

where  $S \in \mathbb{R}^{m \times n}$  is sparse (i.e. having few non-zero entries) and  $L \in \mathbb{R}^{m \times n}$  is low-rank. Under mild assumptions on  $L$  and  $S$  it is possible to recover them via

$$(L^*, S^*) = \arg \min_{\text{rk } L \leq k} \|S\|_0, \text{ s.t. } X = L + S. \quad (7)$$

A Robust PCA method is proposed in [6] that performs a convex relaxation of (7) employing an  $\ell_1$ -penalized outlier matrix  $S$  and minimization of  $\|L\|_*$ , which denotes the nuclear norm. Under specific circumstances this convex method is able to recover the low-rank component exactly. However, only a whole batch of data samples can be processed and the proposed solvers do not achieve realtime performance. The authors of *GoDec* [25] report a significantly faster processing time, which is achieved by using random projections. The method is robust against additive Gaussian noise, but it requires an estimate on the cardinality of the sparse component.

A different way of searching for a low-rank approximation of given data is to employ the so-called Grassmannian, which is the manifold of fixed-dimensional subspaces [3], [17]. In [2] it is shown how the Grassmannian can be exploited for online subspace tracking, i.e. analyzing data sample-wise and constantly adapting its low-rank approximation using a one-step gradient descent. The authors furthermore demonstrate that subspaces can be reconstructed even from highly subsampled data if the upper bound on the desired rank is very low compared to the dimension of the data, which is in the same spirit as [23]. Finally, the *GRASTA* method [15] robustifies subspace tracking using an  $\ell_1$ -norm cost function and achieves close to realtime performance on an online background subtraction task.

Approximating the impractical  $\ell_0$ -norm by the  $\ell_1$ -norm offers the advantage of obtaining a convex problem with a guaranteed globally optimal solution. However, it is known that other measures such as an  $\ell_p$ -norm offer a better approximation of the  $\ell_0$ -norm, cf. [9]. In [12] a way is shown how these kind of  $\ell_0$ -surrogates can be incorporated in an alternating minimization framework for robust subspace estimation and tracking. Numerical results and online background subtraction experiments indicate that using a smoothed  $\ell_p$ -norm sparsifying function increases the robustness of such kind of methods even further. This paper builds on the results of both [12] and [15], and we present a realtime robust online subspace tracking method based on alternating minimization of a smoothed  $\ell_p$ -norm sparsifying function

on manifolds using one-step gradient and conjugate gradient descent.

### 3 The pROST algorithm

As with the classic PCA based models in Section 2.1, we assume that an image  $x \in \mathbb{R}^m$  is generated by a background subspace model with the addition of Gaussian noise  $\epsilon \in \mathbb{R}^m$  and a sparse outlier vector  $s \in \mathbb{R}^m$  which represents the foreground in the scene, i.e.

$$x = Uy + s + \epsilon. \quad (8)$$

Our goals are (i) to robustly recover this background subspace  $U$  from training data containing foreground objects (i.e.  $s \neq 0$ ), (ii) to robustly fit the model to unknown video frames in order to determine the foreground  $s$ , and (iii) to track any changes to the background subspace of a scene.

#### 3.1 Weighted smoothed $\ell_p$ -norm cost function

In [12] it has been shown that smoothed non-convex sparsity measures allow the reconstruction of subspaces from corrupted data in cases where other methods fail. Thus, we construct the cost function based on the smoothed  $\ell_p$ -norm as

$$h_\mu : \mathbb{R}^m \rightarrow \mathbb{R}^+, x \mapsto \sum_{i=1}^m (x_i^2 + \mu)^{\frac{p}{2}}, \quad 0 < p < 1, \quad (9)$$

where  $\mu > 0$  is a smoothing parameter. This particular  $\ell_0$ -surrogate serves as an arbitrary example, for other possible functions see [12]. Even though using (9) leads to a non-convex optimization problem, in practice it is good-natured and can be optimized locally by standard methods.

The pROST algorithm is designed with background subtraction for video streams in mind, and thus we can further tailor the cost function to this setting. In video data it is sensible to assume that spatial and temporal proximity of pixels entail identical semantics. In other words, corresponding pixels in consecutive frames are likely to have the same label. This knowledge can be used to further increase the robustness of the residual cost. The idea is to reduce the contribution of labeled foreground pixels to the overall penalty by introducing additional pixel weights  $w_i \in \mathbb{R}^+$ , whose magnitudes depend on the labels assigned to the pixels in the previous frame. If the pixel was previously labeled a foreground pixel and is therefore likely to remain an outlier in the current frame, the weighting should be small to avoid foreground objects compromising the background. In the reverse case, if the pixel was labeled a background pixel before the weight should be equal to one to allow for model maintenance. In this way the algorithm avoids erroneously

fitting the background model to already known foreground objects and it can focus on fitting the background model to the scene background instead. This extension to the cost function does not only ease bootstrapping from corrupted training data, but it also overcomes the reported difficulties of PCA methods with large foreground objects [4].

We incorporate pixel-weighting by defining the weighted smoothed  $\ell_p$ -norm cost function

$$h_{\mu,w} : \mathbb{R}^m \rightarrow \mathbb{R}^+, x \mapsto \sum_{i=1}^m w_i (x_i^2 + \mu)^{\frac{p}{2}}, \quad 0 < p < 1 \quad (10)$$

and the eventual cost function to be minimized in pROST is

$$h_{\mu,w}(x - Uy). \quad (11)$$

#### 3.2 Optimization on the Grassmannian

The topics of optimization on the Stiefel manifold and the Grassmannian are covered in great detail in [1] and [7]. Here we only recall the most important results and apply them to our specific problem.

In Section 2.1 we define  $U$  to be an element of the so-called Stiefel manifold  $\text{St}_{k,m}$ . However, optimizing over the entire set  $\text{St}_{k,m}$  is not necessary, because whenever  $(U, y)$  is a reasonable solution then so is  $(UQ, Q^\top y)$  for

$$Q \in O(k) := \{Q \in \mathbb{R}^{k \times k} \mid Q^\top Q = I_k\}, \quad (12)$$

where  $O(k)$  denotes the set of  $k \times k$ -dimensional orthogonal matrices. In other words, we are only interested in the subspace spanned by the columns of  $U$ , and not in a particular basis of that subspace, so the search space can be reduced. To that end we employ the well known Grassmannian, which is defined as the quotient manifold

$$\text{Gr}_{k,m} := \text{St}_{k,m} / O(k), \quad (13)$$

with the equivalence relation  $U \sim \tilde{U}$  if and only if there exists a  $Q \in O(k)$  such that  $\tilde{U} = UQ$ . We denote the equivalence class for some representative  $U \in \text{St}_{k,m}$  by

$$[U] = \{\tilde{U} \in \text{St}_{k,m} \mid \tilde{U} \sim U\} \in \text{Gr}_{k,m}. \quad (14)$$

Note that the class  $[U]$  does not have a matrix representation in  $\mathbb{R}^{m \times k}$ . So whenever we store  $[U]$ , we will do that by using one (arbitrary) class representative. In contrast, the Stiefel manifold has a unique matrix representation, as does its tangent space, which is given by

$$T_U \text{St}_{k,m} := \{B \in \mathbb{R}^{m \times k} \mid B^\top U + U^\top B = 0\}. \quad (15)$$

Due to the quotient geometry of  $\text{Gr}_{k,m}$ , the spaces  $T_U \text{St}_{k,m}$  and  $T_{UQ} \text{St}_{k,m}$  share one subspace independently of  $Q$ , which

we identify with the tangent space of the Grassmannian, namely

$$T_{[U]} \text{Gr}_{k,m} := \{B \in \mathbb{R}^{m \times k} \mid U^\top B = 0\}. \quad (16)$$

Optimization on Riemannian manifolds like the Grassmannian is done by moving along tangent space directions on geodesics of the manifold. In order to do this, it is necessary to first project the ambient space gradient

$$\text{grad}_U = \nabla_U h_{\mu,w}(x - Uy) \quad (17)$$

onto the tangent space at  $U$  and then, in the case of minimization, move along the geodesic on the manifold in the opposite direction.

Geodesics are curves that locally minimize the distance between two points on the manifold. For a given tangent direction  $B \in T_{[U]} \text{Gr}_{k,m}$ , the geodesics on the Grassmannian emanating from  $[U]$  in direction  $B$  are given by  $[U(t)]$  with

$$U(t) = (UV \cos(\Sigma t) + \Theta \sin(\Sigma t))V^\top, \quad (18)$$

where  $\Theta \Sigma V^\top = B$  is the compact Singular Value Decomposition of  $B$ , cf. [7].

It is easy to verify that the orthogonal projection of some  $H \in \mathbb{R}^{m \times n}$  onto  $T_{[U]} \text{Gr}_{k,m}$  is given by

$$\pi(H) = (I - UU^\top)H. \quad (19)$$

Using all these components we can formulate a procedure in order to find the background subspace model  $U$ . We propose an alternating approach that iteratively updates  $U$  and  $y$ . The cost function that is minimized is invariant on equivalence classes  $[U]$  when considered with two variables  $(U, y)$ . However, as this is no longer the case when  $y$  is fixed, it is not reasonable to search for an optimal element on  $\text{Gr}_{k,m}$ . Thus, the optimization step for fixed  $y$  will be taken in the direction  $\pi(\text{grad}_U)$  along  $U(t)$  as defined in (18).

### 3.3 An online alternating minimization algorithm

In an online setting the video frames arrive at a certain rate and have to be processed as they arrive. Processing a frame  $x^{(i+1)} \in \mathbb{R}^m$  at time instance  $i + 1$  involves three steps, which are robustly fitting the background model to the frame, updating the background subspace model  $U$  to cope with changes in the background, and segmenting foreground and background for the current frame:

*Step 1:* Refine  $y^{(i)}$  to obtain  $y^{(i+1)}$  via

$$y^{(i+1)} = \arg \min_{y \in \mathbb{R}^k} h_{\mu,w}(x^{(i+1)} - U^{(i)}y) \quad (20)$$

*Step 2:* Take one gradient descent step along  $U(t)$  as defined in (18) to approximate

$$t^* = \arg \min_{t \in \mathbb{R}} h_{\mu,w}(x^{(i+1)} - U(t)y^{(i+1)}) \quad (21)$$

and to obtain the updated subspace  $U^{(i+1)} := U(t^*)$ .

*Step 3:* Identify the outliers or the foreground pixels to obtain the reconstruction cost weighting  $w$  for the next iteration

$$w_i = \begin{cases} 1 & \text{if } |x_i - u_i^\top y| < \delta \\ \omega & \text{else,} \end{cases} \quad (22)$$

where  $\omega$  is the weight for the foreground pixel reconstruction error. In order to be able to slowly incorporate foreground objects into the background,  $\omega$  should be set to a small, but non-zero value.

In Step 1 pROST uses a Conjugate Gradient (CG) algorithm [16] to perform the optimization. Even though one iteration of CG is more expensive than one iteration of a simpler gradient descent algorithm it needs fewer iterations to identify the outliers. Since most of the computational cost per iteration is due to the evaluation of the cost function and computation of the gradient, this actually leads to a more efficient algorithm. Our experiments have shown, that in most cases as little as five CG iterations are sufficient.

In Step 2 pROST takes one gradient decent step on the Grassmannian. This would usually require the costly computation of the full SVD of the projected gradient. In the online setting, however, this can be avoided. The derivative of the cost function with respect to  $U_{i,j}$  is given by

$$\frac{\partial h_{\mu,w}(r)}{\partial U_{i,j}} = -p w_i r_i (r_i^2 + \mu)^{\left(\frac{p}{2}-1\right)} y_j \quad (23)$$

with  $r = x - Uy$ . Using the short notation

$$\eta_i = -p w_i r_i (r_i^2 + \mu)^{\left(\frac{p}{2}-1\right)}, \quad i = 1, \dots, m \quad (24)$$

the projected gradient can be expressed as

$$G_U = \pi(\eta)y^\top. \quad (25)$$

It can easily be verified that  $G_U$  has rank one and its SVD is given by

$$G_U = \sigma_1 s v^\top \quad \text{with} \quad (26)$$

$$s = \frac{\pi(\eta)}{\|\pi(\eta)\|_2}, \quad \sigma_1 = \|\pi(\eta)\|_2 \|y\|_2, \quad v = \frac{y}{\|y\|_2}. \quad (27)$$

Consequently, we are freed of computing the SVD of the search direction at each iteration. This approach has also been taken in GROUSE [2] and GRASTA [15] in order to obtain a fast online gradient decent algorithm for subspace estimation.

### 3.4 Practical issues

**Initialization** In this alternating scheme an initialization for  $U$  has to be provided. We choose to initialize the subspace randomly, which can be performed by computing a reduced QR-decomposition of a random  $m \times k$  matrix. This means that pROST does not use a separate batch initialization phase. It is fully capable of recovering subspaces from video data corrupted with foreground objects. The background subspace is learned one frame at a time while continually reducing the step size from  $t_{init}$  to  $t_{min}$ . The former should be chosen quite high ( $t_{init} \in [10^{-4}, 1]$ ) to facilitate quick initialization, while the latter should be chosen quite low to avoid trailing ghost images of moving foreground objects ( $t_{min} \in [10^{-7}, 10^{-4}]$ ).

The step size for the subspace updates in each iteration are defined by the step-size rule

$$t = \max\{e^{-\tau i} t_{init}, t_{min}\}, \quad (28)$$

where  $i$  is the iteration and  $\tau$  is a parameter controlling the shrinkage rate for the step size reduction. Whenever an initialization phase is defined by an exact number of frames  $i_{init}$ , the parameter  $\tau$  can be calculated as

$$\tau = \frac{-\log\left(\frac{t_{min}}{t_{init}}\right)}{i_{init}}. \quad (29)$$

**Pre- and post-processing** Firstly, the running average of the image data is maintained during the initialization phase and subtracted from each frame before pROST is applied. This means the background subspace has to capture only the dynamic aspects of the scene. Secondly, the images are normalized by dividing the intensity values by the sample standard deviation over all pixels in the initialization phase. In our experiments we observe that this kind of preprocessing is highly beneficial for capturing the scene dynamics. To achieve fast and uniform processing we re-sample all videos to a size of  $160 \times 120$ .

Apart from the thresholding operation, we also apply a  $3 \times 3$  median filter to the foreground segmentation mask  $F$  to fill small holes and to get rid of small clusters of erroneously labeled pixels.

**Color images** If colored video is available, it is clearly advantageous to use the information provided in the color channels for segmentation. We represent a colored vectorized video frame of size  $m$  by a vector

$$x = \begin{bmatrix} x_R \\ x_G \\ x_B \end{bmatrix} \in \mathbb{R}^{3m}, \quad (30)$$

where the  $i$ -th entry of  $x_R, x_G, x_B \in \mathbb{R}^m$  is given by the respective channel value at pixel  $i$ . Accordingly, the background subspace is modeled by

$$U = \begin{bmatrix} U_R \\ U_G \\ U_B \end{bmatrix} \in \text{St}_{k,3m}. \quad (31)$$

The pixel  $i$  is classified as foreground if the difference between the reconstructed background of either of the channels is large enough, i.e. if

$$\max\{|x_{R,i} - u_{R,i}^\top y|, |x_{G,i} - u_{G,i}^\top y|, |x_{B,i} - u_{B,i}^\top y|\} \geq \delta. \quad (32)$$

Here,  $u_{R,i}, u_{G,i}, u_{B,i}$  denote the respective rows of  $U$ .

## 4 Implementation

In order to achieve realtime performance we have implemented pROST on a GPU. More precisely, the preprocessing and all steps of pROST are implemented on the GPU, whereas the median filtering operation in the post-processing stage runs on the CPU. For transferring the images to the GPU we use pinned host memory.

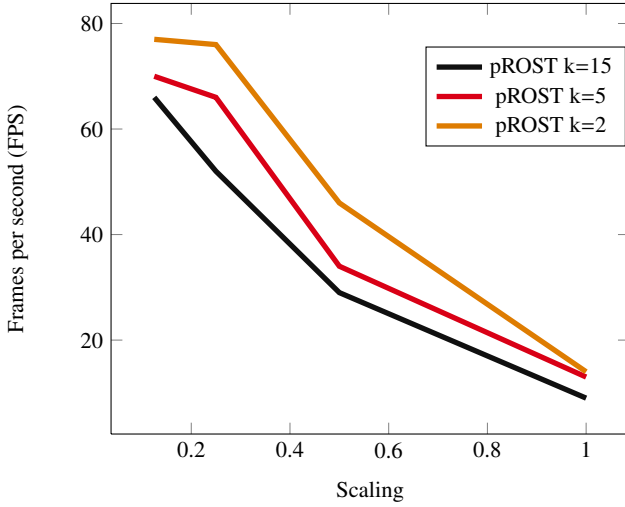
One of the strengths of pROST is its simplicity. Since most of the operations involved are matrix operations, pROST can be parallelized very efficiently on a GPU using C++, CUDA and the highly optimized CUBLAS library for linear algebra operations on the GPU. Step 2 of pROST, for example, can be implemented with as little as four General Matrix Multiply (GEMM) operations and only takes about 5 ms for a subspace dimension of  $k = 15$  and an image resolution of  $160 \times 120$  on a Nvidia GTX 660 GPU.

For images in this resolution and a subspace dimension of 15, more than 50% of the computation time is spent on matrix multiplications, further 30% on basic operations like matrix addition, element-wise multiplication and evaluating the cost function. In order to optimize the implementation even further, we have taken great care to reduce the required number of matrix multiplications and order them in such a way as to reduce the overall complexity and memory requirements. Evaluating the cost function involves a parallel reduction, which has been implemented following the scheme presented in [13].

## 5 Evaluation

Our goal for the evaluation is twofold. Firstly, we rank the pROST method for background subtraction among competing methods on a widely-known benchmark. Secondly, we show that using the weighted smoothed  $\ell_p$ -norm instead of the  $\ell_1$ -norm leads to superior results for background subtraction by comparing our method to GRSTA [15], a state-of-the-art representative of online Robust PCA. As mentioned

**Fig. 1** Performance of the GPU implementation of pROST for several image scalings and subspace dimensions. Scaling is relative to  $320 \times 240$  images.



in Section 1 we conduct all experiments on the *changedetection.net* dataset, and apart from discussing the results here we will also publish the results on the project website to allow a quick and easy comparison with different approaches like GMM-based or non-parametric methods. As the benchmark requires a static configuration for all scenarios we fix all parameters for a first overview and discuss their particular influence afterwards in a more detailed investigation.

### 5.1 Performance on the *changedetection.net* benchmark

The *changedetection.net* dataset [10] consists of six categories of videos and provides ground truth for each frame. The ground truth contains information about background and foreground objects as well as their boundaries and shadows. For some of the videos, the segmentation is evaluated only for certain regions of interest (ROI) while for others the whole image is evaluated. In order to produce comparable results, an evaluation tool is provided which computes significant statistical measures for the segmented images. The evaluation starts after a certain number of frames, which can be used for initialization. However, these training samples have the same foreground-background distribution as the ones used for evaluation and can therefore contain foreground objects.

For the benchmark evaluation we select the following parameters, which maximize the overall performance

- $k = 15$  (subspace dimension),
- $\omega = 5 \times 10^{-5}$  (foreground weighting),
- $t_{init} = 5 \times 10^{-3}$  (initial stepsize),
- $t_{min} = 10^{-4}$  (online stepsize),
- $\delta = 0.35$  (threshold),

- $p = 0.25, \mu = \delta^2(1 - p)$  (smoothed  $\ell_p$ -norm parameters).

For each frame we perform a maximum of five CG steps for the optimization of (20).

The detailed results for pROST are listed in Table 1. By varying the threshold parameter  $\delta \in [0.05, 0.6]$  we obtain the ROC curves for all categories, which are displayed in Figure 2.

In order to compare pROST to GRASTA [15] we rely on the *streaming* version of GRASTA whose MATLAB implementation is available for download on the author’s website<sup>2</sup>. This implementation is intended to work with gray scale images, whereas we work with RGB color images. To allow for a fair comparison we have modified GRASTA to work with such images. We use the same subspace dimension as for pROST and down-sample all images to a resolution of  $160 \times 120$ . GRASTA requires an initialization phase in which an initial background model is learned from a batch of training images. To allow the best possible outcome in this phase we use the largest possible set of training images, i.e. all frames at the beginning of the videos that are not evaluated, and use all the pixels in each video frame to learn the subspace. We allow GRASTA to take three passes over the data, which means that it encounters each video frame three times as often as pROST during the initialization process. We rely on the default parameters of the MATLAB implementation except for the detection threshold and the percentage of pixels used for updating the subspace during the tracking stage. The demo implementation suggests to use 10% of the pixels, while we use 25% of the pixels. The reason for not using all available pixels is that GRASTA is explicitly designed for reconstructing subspaces from incomplete information. In all experiments we have observed that indeed, GRASTA’s performance does not deteriorate markedly if the data is subsampled as the authors describe. We use a value of 0.2 for thresholding in GRASTA, which is twice as high as the threshold suggested by the authors. This choice is motivated by the fact, that the backgrounds in the *changedetection.net* benchmark are highly dynamic and a lower threshold would lead to excessive amounts of false positives. The obtained segmentation masks are post-processed by applying a  $3 \times 3$  median filter. All benchmark results for GRASTA are listed in Table 2.

### 5.2 Discussion

The pROST method excels when the camera is jittering and ranks first in this category by a large margin. In the other categories the method ranks mid-field. It is important to note, however, that it is possible to achieve better performance in the other categories by tuning the parameters individually (see Section 5.3).

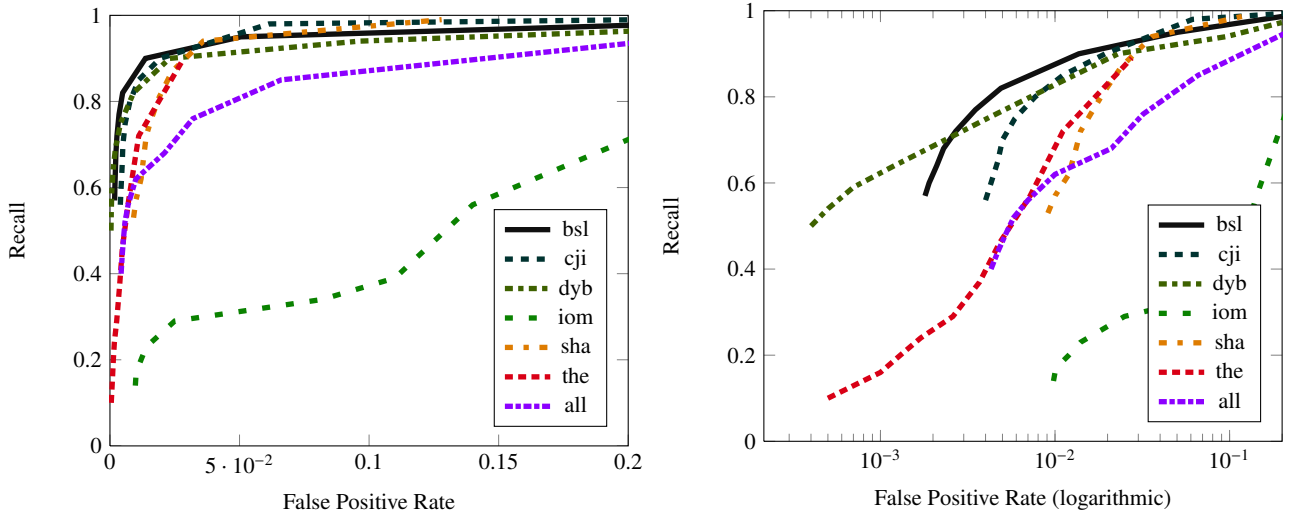
<sup>2</sup> <https://sites.google.com/site/hejunzz/grasta>

**Table 1** Per category results for pROST in the *changedetection.net* benchmark

Category	Recall	Specificity	FPR	FNR	PWC	Precision	FMeasure
baseline	0.801	0.9941	0.0059	0.199	1.28	0.805	0.799
camera jitter	0.770	0.9925	0.0075	0.230	1.56	0.825	0.792
dynamic background	0.743	0.9945	0.0055	0.257	0.73	0.566	0.595
intermittent object motion	0.540	0.9137	0.0863	0.460	9.71	0.488	0.419
shadow	0.754	0.9798	0.0202	0.256	2.85	0.671	0.706
thermal	0.497	0.9920	0.0080	0.503	2.97	0.756	0.584
overall	0.684	0.9778	0.0222	0.316	3.18	0.685	0.650

**Table 2** Per category results for GRASTA in the *changedetection.net* benchmark

Category	Recall	Specificity	FPR	FNR	PWC	Precision	FMeasure
baseline	0.609	0.9926	0.0074	0.391	2.13	0.740	0.664
camera jitter	0.622	0.9282	0.0718	0.378	8.36	0.354	0.434
dynamic background	0.701	0.9760	0.0240	0.299	2.61	0.262	0.355
intermittent object motion	0.311	0.9842	0.0158	0.689	6.32	0.515	0.359
shadow	0.608	0.9554	0.0446	0.392	6.09	0.536	0.529
thermal	0.344	0.9851	0.0149	0.656	6.13	0.726	0.428
overall	0.533	0.9702	0.0298	0.467	5.27	0.522	0.461

**Fig. 2** ROC curves for pROST on the *changedetection.net* benchmark (bsl: baseline, cji: camera jitter, dyb: dynamicBackground, iom: intermittentObjectMotion, sha: shadow, the: thermal, all: overall)

A strength of pROST is clearly dealing with fast variations in the background like camera jitter and scenes with quickly moving foreground objects. The outstanding performance achieved in the *camera jitter* category, which mostly requires the initialization from video heavily corrupted with outliers, shows that the method can bootstrap in very difficult situations. pROST is also capable of dealing with gradual lighting changes. The comparison to GRASTA shows that pROST's performance in the *camera jitter* category is not a general feature of PCA models, but rather a combined result of the cost function, the optimization methods and the introduced foreground pixel weighting.

Situations in which the algorithm fails include the relocation of background objects, for which the performance in the *intermittent object motion* category is a clear indicator. These problems can be alleviated to some degree by ad-

justing the parameters to control the speed of background adaptation. The underlying problem, however, is that the algorithm must adapt to some changes faster than to others. When an object starts to move that was formerly a part of the background, the newly revealed background, which will now be labeled as foreground, has to be integrated into the model as quickly as possible. At the same time, the moving object has to remain in the foreground, even when it stops moving and becomes stationary. When a foreground object becomes stationary it has the same spatio-temporal properties as the newly revealed background, and consequently our algorithm will treat them equally. The demands are therefore conflicting. We argue that the background subtraction algorithm presented here is not especially designed for this task, but through the individual weighting coefficients included in the cost function, pROST could solve this prob-



lem in principle. An unresolved problem remains the occurrence of camouflaging. Besides the median filtering in the post-processing step the algorithm has no means of exploiting spatial correlation of pixel labels within a frame. It has to rely solely on color or intensity differences to make segmentation decisions at pixel level, which makes it inapt to cope with this phenomenon.

### 5.3 Choice of parameters

In the following we explain the most important parameters, their influence on the performance and behavior of pROST and how we set them for the evaluation. The effect of different parameter settings are illustrated by results obtained in the *changedetection.net* benchmark.

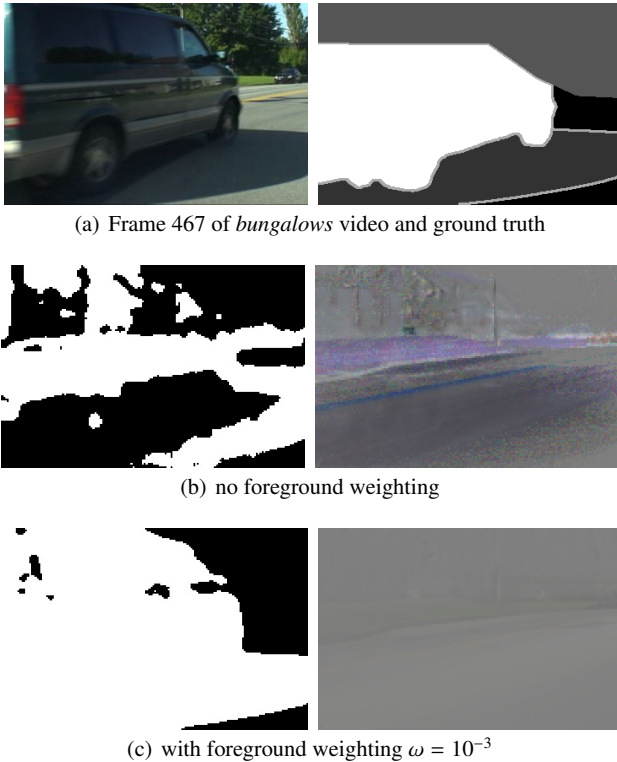
**Subspace dimension** The subspace dimension  $k$  defined by the Stiefel dimension (1) can have considerable impact on the computational complexity of the algorithm, but also on other performance measures. Choosing an overly large value for  $k$  leads to excessive computational complexity while not increasing performance. Low-complexity backgrounds like those in the *baseline* category can be represented with a dimension as low as  $k = 2$ , while the representation of highly dynamic backgrounds like in the *camera jitter* category can

benefit from a higher limit on the subspace dimension. In Figure 3 the results for three different choices of  $k$  and the resulting background and segmentation for frame # 1381 of *traffic* can be seen. To get a more detailed impression of how the background is represented by pROST, we provide some further insight in Figure 4. Notice that the background contains only the dynamic aspects of the scene as the running average is subtracted from each video frame. With growing subspace dimension finer details of the background can be captured, but the model is also getting more flexible in areas that do not require this flexibility. This causes integration of foreground objects into the background. We furthermore observed that there is a strong relationship between image sizes and required subspace dimension. For images of size  $160 \times 120$  a subspace dimension of 10 to 15 is sufficient, even for the highly dynamic *camera jitter* backgrounds. For images of size  $320 \times 240$  the *camera jitter* category requires much higher-dimensional subspaces and also a longer initialization. This problem can be mitigated by reducing the information content in the image, for example by band-limiting the image with a Gaussian blur filter. Another approach is downsampling the image before the processing and upsampling the segmentation mask for the evaluation. This is clearly preferable, because it also reduces computational complexity.

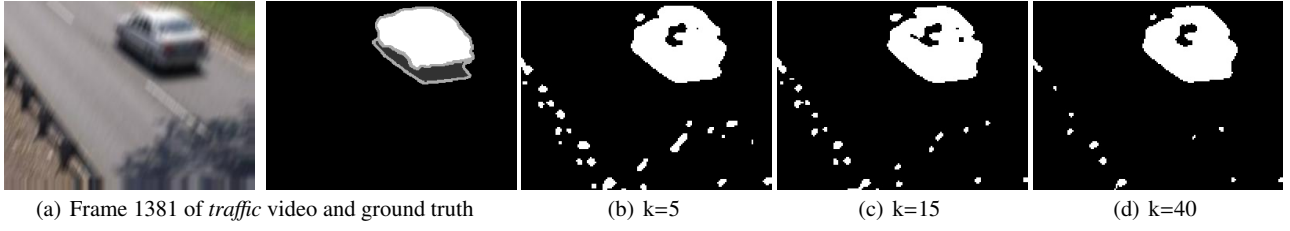
**Foreground weighting** The foreground weighting parameter  $\omega$  from (22) has a large effect on the algorithm's bootstrapping capability, how it deals with highly dynamic complex backgrounds and robustness to large foreground objects. In some scenarios like e.g. the *bungalows* video in the *shadow* category foreground weighting is a crucial component for recovering a background model that is not corrupted by foreground objects. Figure 5 showcases this effect. We found that foreground weighting allows for larger step sizes  $t_{min}$  without producing ghost images.

**Step size** The interplay between foreground weighting and different choices for the step size is displayed in Figure 6. Without any weighting the dynamic elements can be compensated using a large step size, but large foreground objects are incorporated too quickly into the background, which leads to reconstruction failure in some cases. A very small step size prohibits the formation of ghost images, but makes it impossible to adapt to the dynamic elements and lighting changes. Finally, combining a large step size and foreground weighting solves both problems.

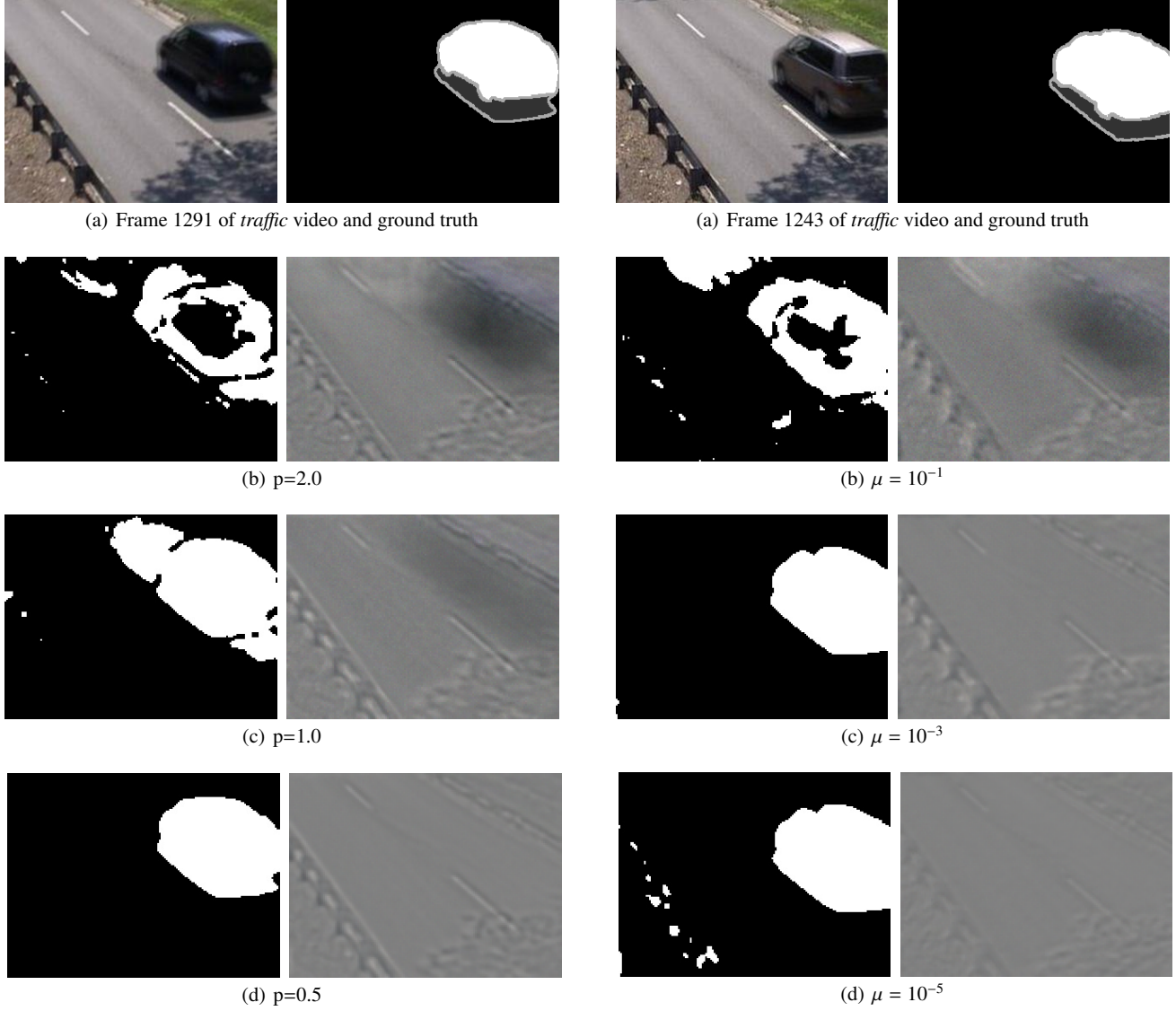
**Cost function parameters** In contrast to the  $\ell_1$ -norm, the choice of  $p$  and  $\mu$  in the smoothed  $\ell_p$ -norm (9) offers control over the degree of robustness to outliers in the data. To analyze this effect, we extend the definition of the smoothed  $\ell_p$ -norm cost function to the case  $p \geq 1$ . Figure 7 illustrates



**Fig. 5** Effects of foreground weighting on dynamic backgrounds with extremely large foreground objects ( $k = 15$ ,  $p = 0.25$ ,  $\mu = 10^{-3}$ )



**Fig. 3** Effects of subspace dimensionality in jittery videos ( $p = 1.0, \mu = 10^{-3}, \omega = 1.0$ )



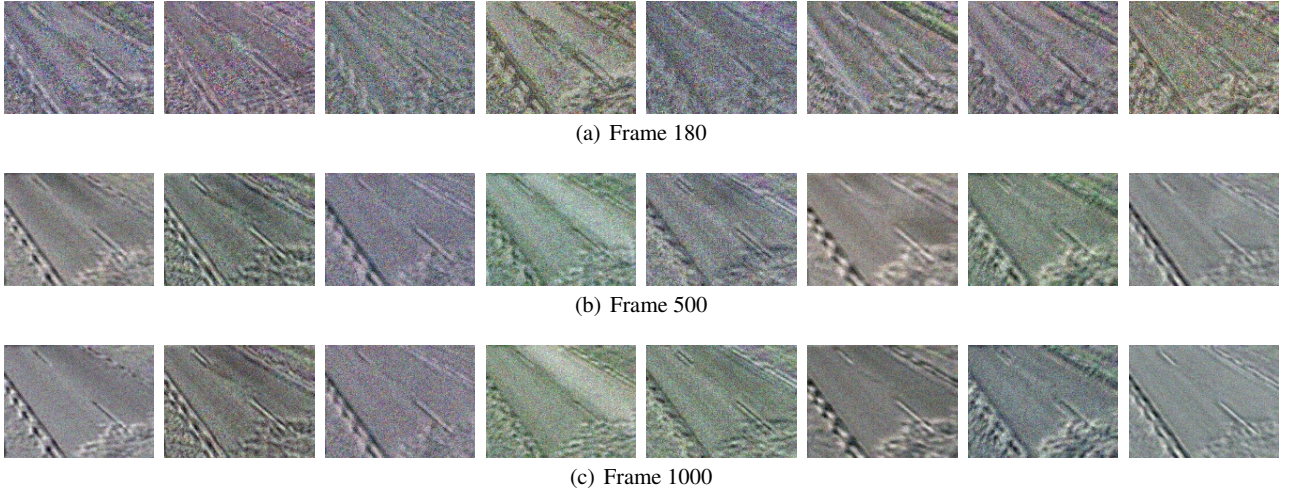
**Fig. 7** Effects of  $\ell_p$ -norm parameter  $p$  on background model learning and fitting in jittery videos ( $k = 15, \mu = 10^{-3}, \omega = 1.0$ )

**Fig. 8** Effects of  $\ell_p$ -norm parameter  $\mu$  on background model learning and fitting in jittery videos ( $k = 15, p = 0.5, \omega = 1.0$ )

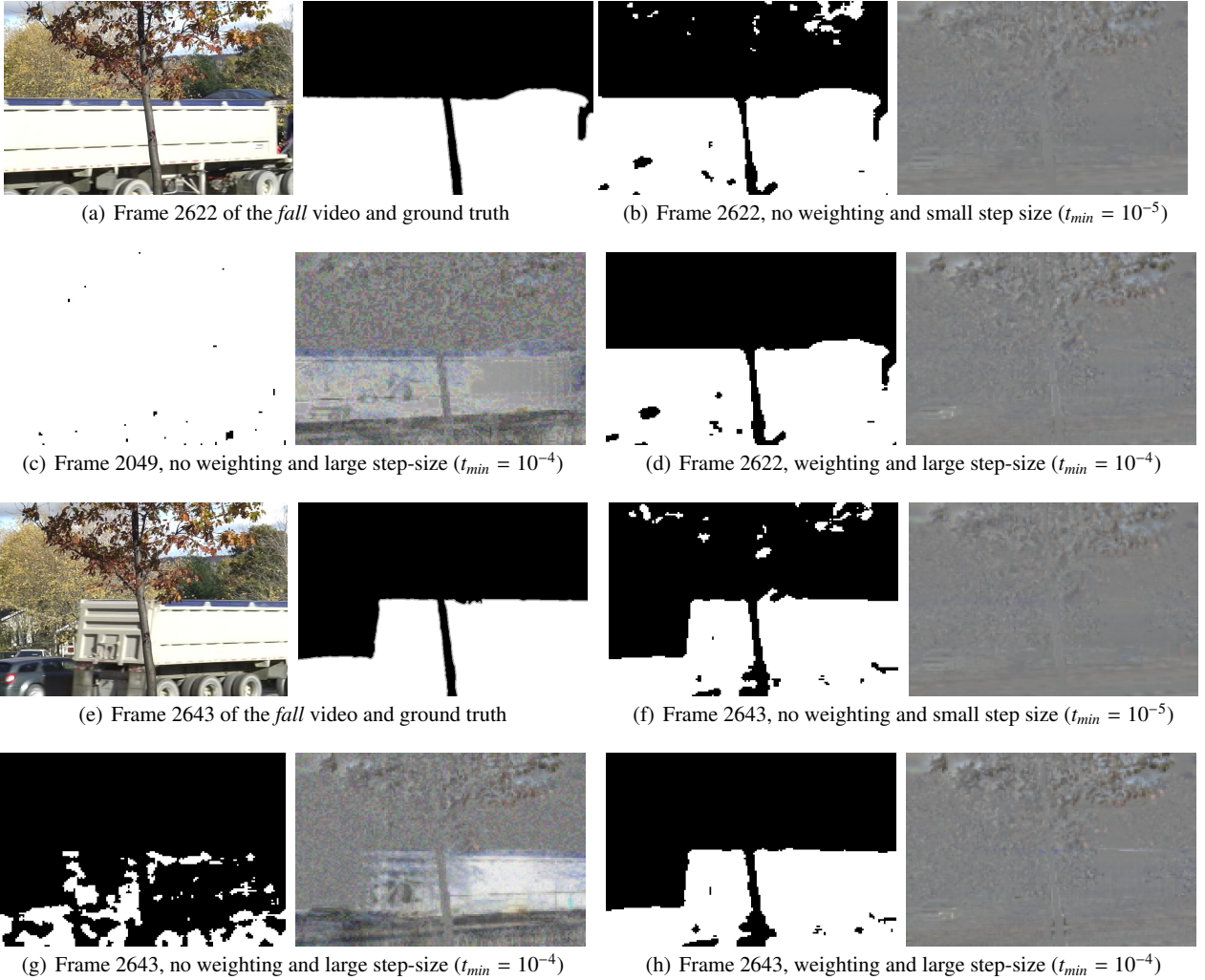
that lowering the parameter  $p$  can reduce the rate of incorporating foreground objects into the background.

As discussed in Section 2, Robust PCA algorithms aim for accurately recovering the assumed low-rank component  $L$  of a data matrix  $X$ . However, when a decision is based on

thresholding of  $S = X - L$ , a certain degree of reconstruction accuracy is sufficient and further decreasing the reconstruction error after it has fallen below the threshold does not add to the performance, but unnecessarily requires computational resources. In this sense, an estimate of  $L$  should be



**Fig. 4** Elements of the evolving 15 dimensional subspace for the *traffic* dataset.



**Fig. 6** Effects of foreground weighting on bootstrapping with extremely large foreground objects ( $k = 15$ ,  $p = 0.5$ ,  $\mu = 10^{-3}$ )



considered sufficiently close to the true  $L$  if it does not produce a false positive. To reflect this idea in the cost function, we require that starting at the threshold  $\delta$  the partial derivative of the cost function with respect to the reconstruction error becomes smaller as we approach zero and becomes larger as we approach the threshold from above. This translates to the constraint

$$\left. \frac{\partial^2 h_\mu}{\partial r_i^2} \right|_{r_i=\delta} = 0, \text{ for fixed } p < 1. \quad (33)$$

A quick calculation reveals that

$$\mu = \delta^2(1 - p). \quad (34)$$

meets this constraint. To back-up that this coupling between  $\mu$  and  $\delta$  indeed leads to near-optimal smoothing, we conduct an experiment in which we evaluate a number of smoothing parameters for two different thresholds on the *traffic* and *badminton* videos. The results in Figure 5.3 confirm our theoretical analysis. Furthermore, the performance degrades markedly if a very small smoothing parameter is chosen.

## 6 Conclusion

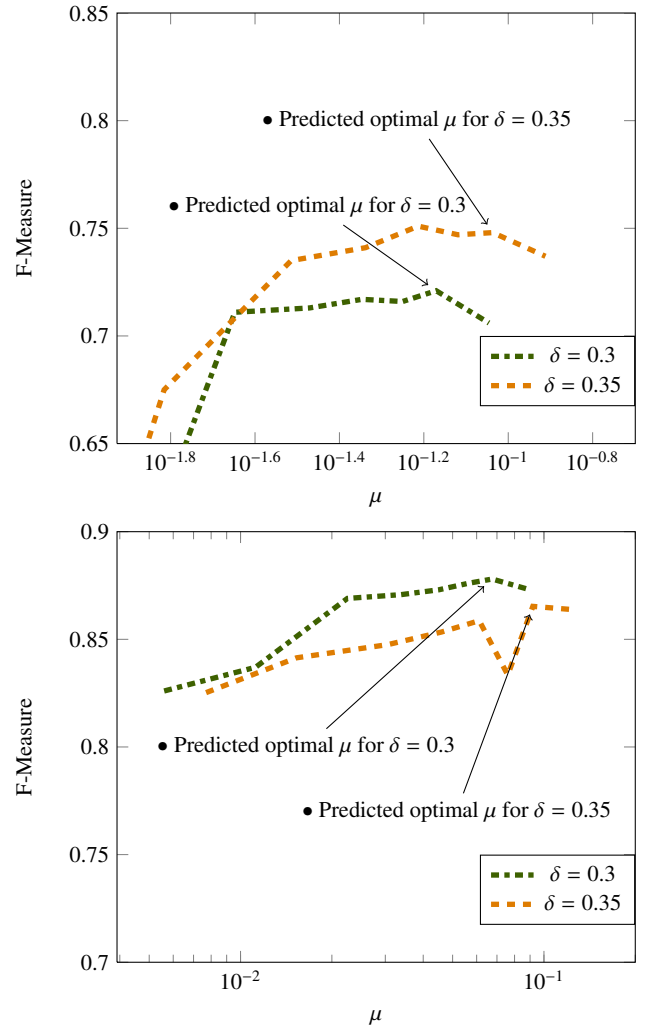
We have presented a novel subspace tracking algorithm which combines concepts of previous methods and introduces a novel weighted robust cost function tailored to the task of background modelling and foreground segmentation from video data. The method is implemented on a GPU, achieves frame rates between 30 and 45 FPS on images in a resolution of  $160 \times 120$  and is thus real-time capable. One of the noteworthy features of the method is that it does not need a batch initialization phase, but learns the background model from corrupted streaming video. This has the advantage that no camera frames need to be stored at any time during operation.

pROST should be considered a basic building block for larger, more sophisticated systems for background subtraction. Future work will include the extension with a shadow detection mechanism and more sophisticated pre- and post-processing techniques.

We have evaluated the algorithm on the *changedetection.net* benchmark and show that it outperforms the conceptually similar GRASTA algorithm in many categories. Our method is particularly suitable for videos recorded by highly unstable cameras, ranking first in this category by a large margin, and it can thus be considered an advancement in this research area.

## References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ (2008)



**Fig. 9** Evaluation of  $\mu$ -heuristic 34 for *traffic* (top) and *badminton* (bottom)

2. Balzano, L., Nowak, R., Recht, B.: Online identification and tracking of subspaces from highly incomplete information. In: Allerton Conference on Communication, Control, and Computing, pp. 704–711 (2010)
3. Boumal, N., Absil, P.A.: RTRMC: A Riemannian trust-region method for low-rank matrix completion. In: Advances in Neural Information Processing Systems, pp. 406–414 (2011)
4. Bouwmans, T.: Subspace learning for background modeling: A survey. *Recent Patents on Computer Science* **2**(3), 223–234 (2009)
5. Brutzer, S., Höferlin, B., Heidemann, G.: Evaluation of background subtraction techniques for video surveillance. In: Computer Vision and Pattern Recognition, pp. 1937–1944. IEEE (2011)
6. Candès, E., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? *Journal of ACM* **58**(3), 1–37 (2011)
7. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications* **20**(2), 303–353 (1998)
8. Elhabian, S.Y., El-Sayed, K.M., Ahmed, S.H.: Moving Object Detection in Spatial Domain using Background Removal Techniques - State-of-Art. *Recent Patents on Computer Science* **1**, 32–34 (2008)

9. Gasso, G., Rakotomamonjy, A., Canu, S.: Recovering Sparse Signals With a Certain Family of Nonconvex Penalties and DC Programming. *Transactions on Signal Processing* **57**(12), 4686–4698 (2009)
10. Goyette, N., Jodoin, P., Porikli, F., Konrad, J., Ishwar, P.: Changedetection.net: A new change detection benchmark dataset. In: *Computer Vision and Pattern Recognition Workshops*, pp. 1–8 (2012)
11. Guyon, C., Bouwmans, T., Zahzah, E.: Robust Principal Component Analysis for Background Subtraction: Systematic Evaluation and Comparative Analysis. In: *Principal Component Analysis*, chap. 12, pp. 223–238. INTECH (2012)
12. Hage, C., Kleinsteuber, M.: Robust PCA and subspace tracking from incomplete observations using  $\ell_0$ -surrogates. Technical Report, Technische Universität München (2012)
13. Harris, M.: Optimizing Parallel Reduction in CUDA. <http://developer.download.nvidia.com/assets/cuda/files/reduction.pdf> (2008)
14. Hassanpour, H., Sedighi, M., Manashty, A.R.: Video frame's background modeling: Reviewing the techniques. *J. Signal and Information Processing* **2**(2), 72–78 (2011)
15. He, J., Balzano, L., Szlam, A.: Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video. In: *Computer Vision and Pattern Recognition*, pp. 1568–1575 (2012)
16. Hestenes, M.R., Stiefel, E.: Methods of Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards* **49**, 409–436 (1952)
17. Keshavan, R.H., Montanari, A.: Matrix completion from noisy entries. *The Journal of Machine Learning Research* **11**, 2057–2078 (2010)
18. Li, L., Huang, W., Gu, I.Y.H., Tian, Q.: Statistical modeling of complex backgrounds for foreground object detection. *Transactions on Image Processing* **13**(11), 1459–1472 (2004)
19. Li, Y.: On incremental and robust subspace learning. *Pattern Recognition* **37**, 1509–1518 (2004)
20. Oliver, N., Rosario, B., Pentland, A.: A bayesian computer vision system for modeling human interactions. *Transactions on Pattern Analysis and Machine Intelligence* **22**(8), 831–843 (2000)
21. Pearson, K.: On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**(6), 559–572 (1901)
22. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: Principles and practice of background maintenance. *International Conference on Computer Vision* **1**, 255–261 (1999)
23. Waters, A., Sankaranarayanan, A.C., Baraniuk, R.G.: SpaRCS: Recovering Low-Rank and Sparse Matrices from Compressive Measurements. In: *Advances in Neural Information Processing Systems* (2011)
24. Xu, Z., Shi, P., Gu, I.Y.H.: An eigenbackground subtraction method using recursive error compensation. In: Y. Zhuang, S. Yang, Y. Rui, Q. He (eds.) *PCM, Lecture Notes in Computer Science*, vol. 4261, pp. 779–787 (2006)
25. Zhou, T., Tao, D.: GoDec: Randomized low-rank & sparse matrix decomposition in noisy case. In: *International Conference on Machine Learning*, pp. 33–40 (2011)